UN OUTIL POUR L'ENSEIGNEMENT DE L'ANALYSE STATISTIQUE TEXTUELLE ET LES SYSTÈMES DE GESTION DE BASES DE DONNÉES

Serge SABOURIN[1], Abdellah QANNARI[2] et Laurence REBOUL[3]

TITLE

A tool for textual analysis teaching and Data Base Management

RÉSUMÉ

Dans un souci de transversalité entre les enseignements de statistique et d'informatique dispensés dans nos formations, nous avons élaboré un logiciel pour la préparation et les analyses de base d'un corpus textuel. Cet outil permet d'accompagner nos étudiants dans la mise en application de leurs connaissances à la fois en analyse statistique textuelle et en gestion de bases de données.

Mots-clés : *analyse textuelle, SGBD, Excel, Access.*

ABSTRACT

For the sake of transversality between informatics and statistics teaching, we elaborated a software for the preparation and primary analysis of a textual corpus. This tool assists our students in the application of their knowledge in textual statistical analysis and Data Base Management.

Keywords: *textual analysis, DBMS, Excel, Access.*

1 Introduction

Les formations professionnalisantes telles que celles que nous dispensons dans le cadre de l'Institut Universitaire de Technologie de STatistique et Informatique Décisionnelle (IUT STID) et de la licence professionnelle Statistiques Commerciales doivent répondre à la nécessité croissante de stocker, gérer et traiter des données de plus en plus volumineuses dans de nombreux champs économiques. Cette finalité mobilise de solides compétences tant en informatique qu'en statistique.

L'outil présenté ici est à visée purement pédagogique et a été créé dans un souci de transversalité entre ces deux types de compétences. Il s'agit d'un logiciel d'analyse textuelle conçu à l'aide des logiciels ACCESS et EXCEL de la suite Microsoft Office, dont la prise en main s'inscrit dans le programme d'enseignement de nos étudiants. Il est composé de deux applications : anarapport.xls et anarapport.mdb qui fonctionnent dès lors que les macros associées ont été autorisées. Complètement modulable par l'utilisateur, ce logiciel participe à l'apprentissage des Systèmes de Gestion de Bases de Données relationnelles tout en permettant aux étudiants de se familiariser avec les méthodes statistiques utilisées en analyse textuelle, dont l'objectif est d'extraire d'un ensemble de textes (discours, questions ouvertes, etc.), communément appelé corpus, l'information la plus pertinente. Ces méthodes ont connu

[1] IUT de Poitiers, Département STID, Niort, serge.sabourin@math.univ-poitiers.fr
[2] IUT de Poitiers, Département STID, Niort, abdellah.qannari@univ-poitiers.fr
[3] Université de la Méditerranée, Marseille, reboul@iml.univ-mrs.fr

un grand essor depuis les années quatre-vingt dans différents domaines tels que les sciences sociales, les sciences de gestion, la stylométrie, la recherche documentaire (voir Gauzente et Peyrat-Guillard (2007), Jeandillou (2006), Lebart et Salem (1994) pour un aperçu de ce champ de la statistique).

En tant qu'instrument d'apprentissage d'ACCESS et de la statistique textuelle, notre outil a été conçu de manière à permettre aux étudiants de modifier des modules existants ou d'en développer de nouveaux en fonction de leurs besoins, au fur et à mesure de leur acquisition des méthodes de la statistique textuelle en cours de statistique. L'outil possède donc, dans sa version de base, des fonctionnalités simples destinées à être améliorées et ne prétend pas rivaliser avec les logiciels spécialisés tels que SPAD.T, ALCESTE, HYPERBASE, PROSPERO, etc. (voir De Saint Pol (2003) pour une description de ces logiciels). Il permet tout au plus de se familiariser avec les outils les plus simples d'exploration statistique d'un corpus textuel : dépouillement du corpus et recoupement de données en utilisant le comptage de mots, la lemmatisation, l'étude des contextes. Toutefois, il possède l'originalité de permettre à l'utilisateur de créer des indices pour quantifier l'adéquation d'un texte à une thématique, ce qui ne semble pas avoir été développé jusqu'alors par les logiciels existants.

2 Présentation de l'outil

Le fonctionnement de l'outil (https://bv.univ-poitiers.fr/access/content/group/8edda960-aa89-448e-aaf9-4a16f9238997/Anatexte/anarapports%20synon4.mdb) requiert deux étapes. Une première étape de préparation du corpus et une seconde consacrée à l'analyse à proprement parler de ce corpus.

La première étape est exécutée sous EXCEL à l'aide de macros VBA et finalisée sous une application ACCESS, qui sera aussi utilisée pour effectuer l'analyse des textes. Les menus de l'application permettant les différents traitements sont des formulaires. Les données textuelles et calculs effectués sur celles-ci sont stockés dans des tables ou des requêtes, dont certaines n'ont pas encore été intégrées dans les formulaires. Ces objets ACCESS sont accessibles en lecture et écriture depuis le menu « Base de données » de l'application.

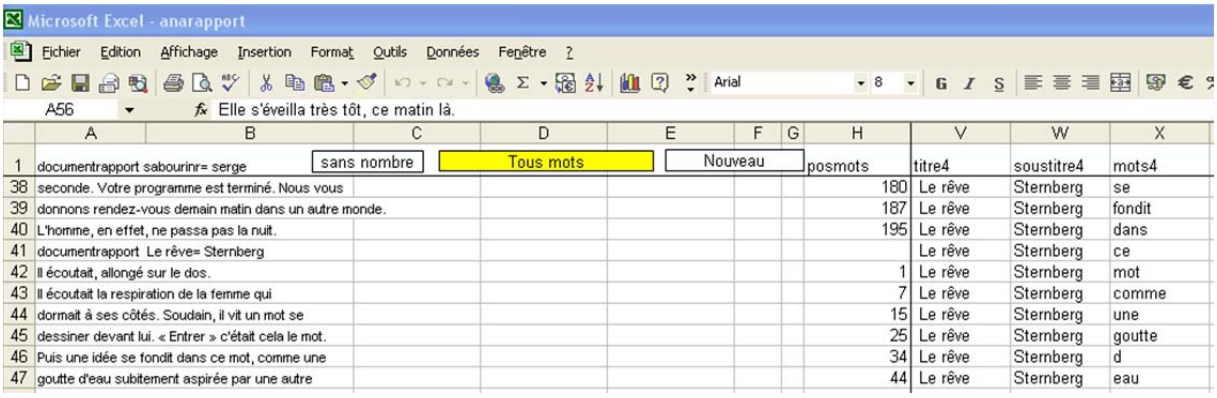
Les deux étapes sont illustrées à partir d'un corpus constitué de trois contes de Sternberg, issus du recueil *Contes glacés* : *Les chats* (329 mots, 23 phrases), *Le communiqué* (203 mots, 15 phrases), *Le rêve* (184 mots, 18 phrases).

2.1 Préparation du corpus

Segmentation et importation des textes

L'objectif premier de cette étape est de transformer les textes à étudier en une structure d'informations requêtables. Chaque nouveau texte à analyser est copié dans le classeur EXCEL nommé anarapport.xls en cliquant sur l'onglet « Nouveau » et en entrant le texte dans la première cellule située juste en dessous du dernier texte entré. Le texte doit au préalable avoir été débarrassé de ses imperfections (fautes d'orthographe, etc.) et potentielles confusions (traitement des sigles séparés par des points tels que S.N.C.F, des abréviations, des verbes suivis d'un pronom personnel tels que peut-il…). La segmentation du texte est effectuée par une macro, qui s'exécute en cliquant sur l'onglet « Tous mots » (l'onglet « sans

nombre » permet d'effectuer une segmentation du texte privé de ses informations numériques). Celle-ci effectue le découpage du texte en phrases, une phrase étant définie comme l'ensemble des mots situés entre deux caractères délimitateurs (qui peuvent être ici le point, les points d'exclamation, d'interrogation ou de suspension). Chaque phrase est débarrassée de ses signes de ponctuation et se présente comme une séquence d'occurrences de mots. La macro affecte à chaque occurrence son rang d'apparition dans le texte (visualisable dans la feuille « Rapports » du classeur) et indique pour chaque phrase sa longueur ainsi que les positions de ses premier et dernier mots dans le texte (visualisables dans la feuille « posphrases »).



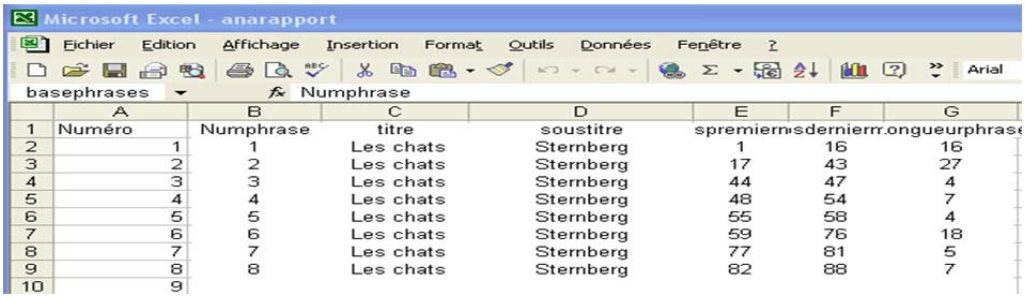
Microsoft Excel - anarapport

Fichier Edition Affichage Insertion Format Outils Données Fenêtre ?

A56 Elle s'éveilla très tôt, ce matin là.

	A	B	C	D	E	F	G	H	V	W	X
1	documentrapport sabourinr= serge		sans nombre	Tous mots	Nouveau			posmots	titre4	soustitre4	mots4
38	seconde. Votre programme est terminé. Nous vous							180	Le rêve	Sternberg	se
39	donnons rendez-vous demain matin dans un autre monde.							187	Le rêve	Sternberg	fondit
40	L'homme, en effet, ne passa pas la nuit.							195	Le rêve	Sternberg	dans
41	documentrapport Le rêve= Sternberg								Le rêve	Sternberg	ce
42	Il écoutait, allongé sur le dos.							1	Le rêve	Sternberg	mot
43	Il écoutait la respiration de la femme qui							7	Le rêve	Sternberg	comme
44	dormait à ses côtés. Soudain, il vit un mot se							15	Le rêve	Sternberg	une
45	dessiner devant lui. « Entrer » c'était cela le mot.							25	Le rêve	Sternberg	goutte
46	Puis une idée se fondit dans ce mot, comme une							34	Le rêve	Sternberg	d
47	goutte d'eau subitement aspirée par une autre							44	Le rêve	Sternberg	eau

FIGURE 1 – Les chats : *Décomposition en mots du conte (extrait) sur la feuille « rapport »*



Microsoft Excel - anarapport

Fichier Edition Affichage Insertion Format Outils Données Fenêtre ?

basephrases Numphrase

	A	B	C	D	E	F	G
1	Numéro	Numphrase	titre	soustitre	spremiern	sdernierm	ongueurphrase
2	1	1	Les chats	Sternberg	1	16	16
3	2	2	Les chats	Sternberg	17	43	27
4	3	3	Les chats	Sternberg	44	47	4
5	4	4	Les chats	Sternberg	48	54	7
6	5	5	Les chats	Sternberg	55	58	4
7	6	6	Les chats	Sternberg	59	76	18
8	7	7	Les chats	Sternberg	77	81	5
9	8	8	Les chats	Sternberg	82	88	7
10	9						

FIGURE 2 – *Synthèse des phrases du conte* Les chats *dans la feuille « posphrases »*

Les fichiers ainsi segmentés doivent ensuite être importés dans les tables « basemots » et « basephrases » de la base de données de l'application ACCESS. Pour ce faire, on ouvre le fichier nommé anarapport.mdb. Un formulaire de menu général « Importation de nouveaux textes » apparaît, qui permet d'importer les nouveaux textes via les onglets « Importer les textes Excel » (resp. « Remplacer les textes par les textes EXCEL »). Les fichiers importés viennent enrichir (resp. remplacer) l'ensemble des textes déjà disponibles dans les tables. Il s'agit dès lors de sélectionner, parmi cet ensemble de textes, ceux sur lesquels portera l'analyse. Pour cela, on doit cliquer sur l'onglet « Aller menu » puis « Suite ».

L'application s'ouvre alors sur un nouveau menu dans lequel les textes préchargés sont présentés avec leur titre et leur nombre d'occurrences sur la sous-fenêtre de droite. Le corpus sur lequel portera l'analyse est sélectionné en cochant les textes que l'on veut inclure puis en validant cette sélection à l'aide du bouton « Valider la sélection des textes ». Le bouton « Sélectionner et valider tous les textes » permet en outre de sélectionner tous les textes importés.

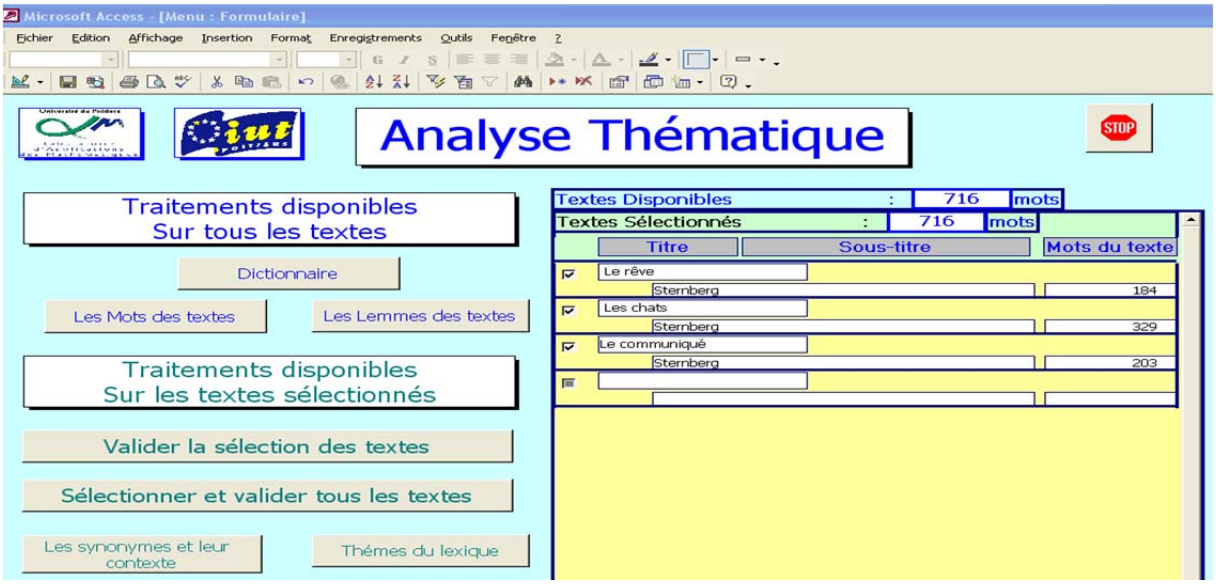


FIGURE 3 – *Menu général « analyse thématique » de l'application access*

Gestion des définitions et des lemmes

Le logiciel possède un dictionnaire de mots dynamique, accessible depuis l'onglet « Dictionnaire ». Il est constitué d'un dictionnaire de base préchargé dans la table « dico » du logiciel, que l'on peut éventuellement enrichir des nouveaux mots contenus dans les textes importés. Le menu permet de faire défiler les mots du dictionnaire ou de rechercher un mot particulier en utilisant les boutons « Suivant », « Précédent » ou la barre de défilement de la fenêtre « Choisir ». La fenêtre « Longueur » indique la longueur du mot. Son rang d'apparition dans le dictionnaire est donné dans la fenêtre « Enr » en bas à gauche de l'écran.

A chaque mot peut être associé s'il y a lieu, son type, son genre, son nombre et un lemme. Ces champs ont été renseignés pour la plupart des mots du dictionnaire préchargé et devraient l'être à terme pour tous. Il est toutefois nécessaire de les préciser pour chaque nouveau mot du corpus rajouté au dictionnaire. La lemmatisation des mots sera en effet utile lors de certaines analyses présentées ultérieurement. Elle consiste à associer à une forme de base ses différentes formes dérivées. La forme de base diffère selon la nature du mot sélectionné, et on ramène généralement :

- les formes verbales à l'infinitif,
- les substantifs au singulier,
- les adjectifs au masculin singulier.

Il est à noter que certaines fonctionnalités de ce menu ne sont pas encore actives et destinées à être développées par les étudiants. En particulier, l'information « Rang dans le dictionnaire », censée indiquer (comme lu dans « Enr ») et stocker la position des mots dans le dictionnaire, pourrait permettre à terme de numériser le corpus afin notamment d'optimiser le stockage en mémoire du texte. Par ailleurs, ce dictionnaire ne permet pas pour l'instant de traiter les homonymes, un mot ne pouvant être entré qu'une seule fois.

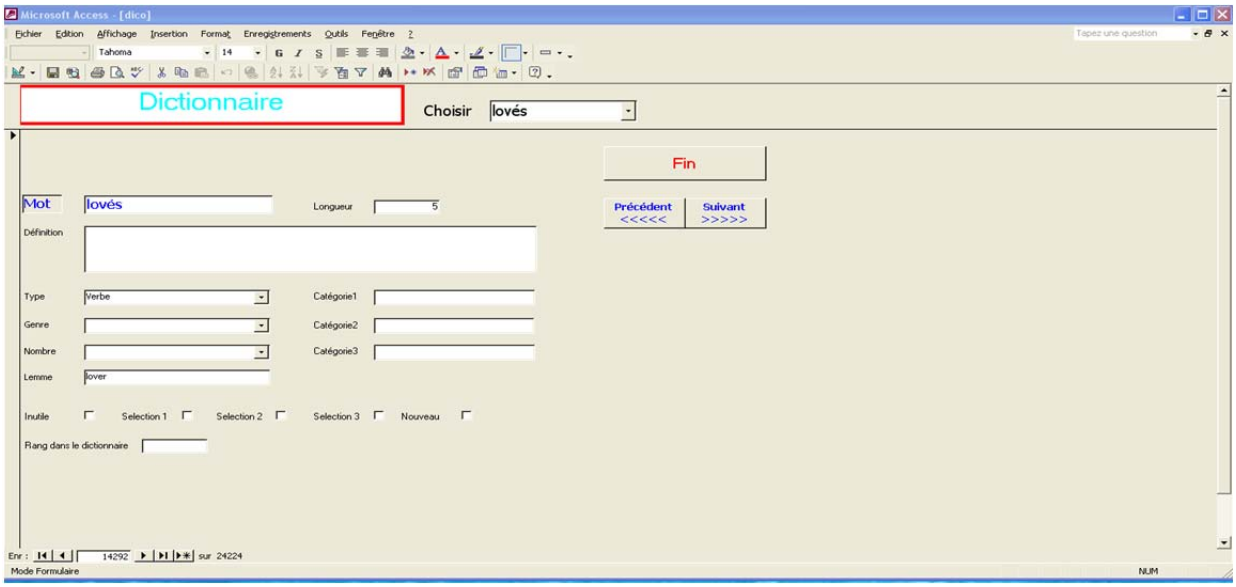


FIGURE 4 – *Aperçu du menu « Dictionnaire »*

2.2 Analyse du corpus

Une fois le dictionnaire à jour, l'utilisateur a la possibilité d'appliquer différents traitements aux textes du corpus.

Etude de la fréquence des formes simples et lemmatisées d'un texte

Les menus « Les mots des textes » et « Les lemmes des textes » permettent respectivement d'obtenir les fréquences des formes simples (mots) ou lemmatisées (lemmes) de n'importe quel texte disponible. Pour l'étude des formes lemmatisées, le module s'appuie sur le travail de lemmatisation effectué lors de la constitution du dictionnaire. Le texte d'intérêt est sélectionné en utilisant la barre de défilement de la fenêtre « Choisir ». Les titre et sous-titre s'inscrivent alors dans la fenêtre de gauche, ainsi que le nombre d'occurrences contenues dans le texte. Le module calcule le nombre d'occurrences et la fréquence d'apparition de toute forme de longueur strictement supérieure au chiffre inscrit dans « Longueur mini », ce paramètre étant ajustable par l'utilisateur (il doit valoir 0 pour inclure toutes les formes du texte). Il indique de plus, au bas de la fenêtre de droite, le nombre de formes différentes repérées.

Dans le texte *Le rêve*, les formes les plus fréquentes sont « elle », « être » et « dans » indiquant que le personnage principal (elle ici) est immergé dans l'univers dans lequel il se trouve dans le texte. D'un point de vue lexical, on trouve 56% de mots différents (nombre de mots/nombre d'occurrences), 36,4% d'hapax (mots n'ayant qu'une seule occurrence), et une fréquence maximale de 26% (nombre d'occurrences du mot le plus fréquent/nombre d'occurrences) qu'il serait intéressant de comparer aux mêmes indicateurs sur un corpus important de textes variés.

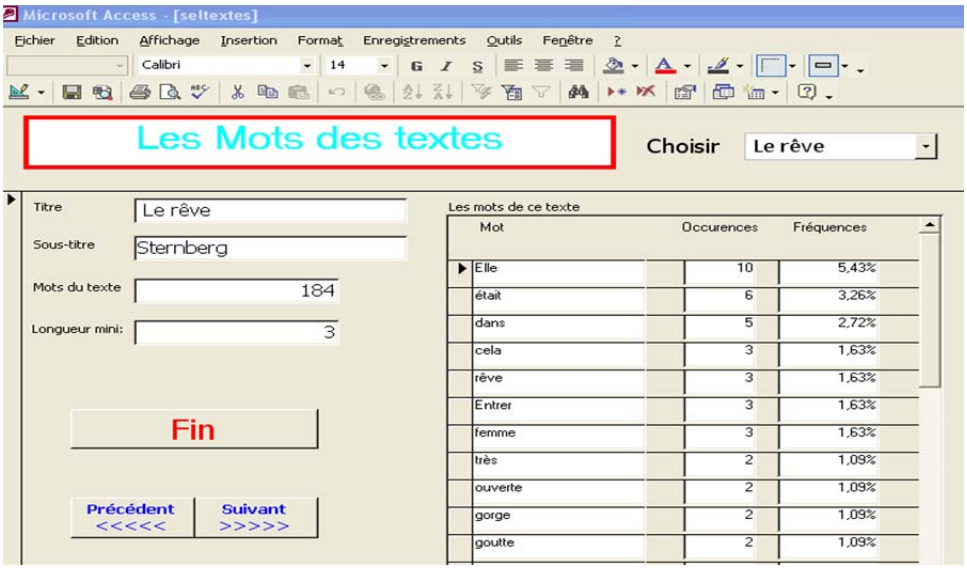


FIGURE 5 – *Les mots du texte* Le rêve

La localisation dans le texte des occurrences d'une même forme n'est pas disponible ici mais sera proposée dans le menu « Les synonymes et leur contexte ». Notons que la plupart des logiciels d'analyse textuelle proposent certaines méthodes graphiques comme le tracer du diagramme de Pareto (graphique représentant la gamme des fréquences des formes d'un texte), des comparaisons de la gamme des fréquences obtenue avec la loi de Zipf. La création de ces fonctionnalités fait partie des développements du logiciel qui sont proposés à nos étudiants. De même, on peut envisager de développer un module permettant des comparaisons simples entre différents textes, telles que celle de leur richesse lexicale, avec la présentation sur la même fenêtre des valeurs des indices de richesse lexicale (calculés plus haut sur le texte *Le rêve)* relatifs à chaque texte.

Notons enfin que le problème de la distinction des homonymes ne peut pas être géré de façon systématique et doit être envisagé au cas par cas par l'utilisateur, en étudiant par exemple les contextes des occurrences de la forme concernée. En effet, d'une part le dictionnaire ne prend pas différents sens d'un même mot, bien que cette possibilité puisse être développée. D'autre part, cette distinction n'est pas envisageable sans module d'analyse grammaticale du corpus, dont le développement dépasse le cadre de cet outil.

Etude des familles de mots et de leurs contextes

Le menu « Les synonymes et leur contexte » vise en premier lieu à créer des familles de mots qui permettront de faire ressortir plus clairement la coloration des textes étudiés. Chaque mot du corpus peut être affecté à une famille identifiée par une étiquette appelée « synonyme », selon des règles d'affectation déterminées par un expert ou l'utilisateur lui-même. Notons que le concept de synonyme recouvre ici un sens plus large que celui qui lui est usuellement donné et s'apparente mieux au concept d'isotopie. En particulier, on peut affecter à un mot donné lui-même, son lemme, ou tout autre mot se rapportant au même champ lexical ou à un même thème. En pratique, le mot à affecter est recherché parmi ceux des textes du corpus en le tapant dans la fenêtre « Choisir » ou à l'aide des boutons « Suivant » ou « Précédent ». Le synonyme est inscrit dans la fenêtre « Synonyme affecté » (par défaut, le module lui affecte son lemme, défini dans le dictionnaire). La fenêtre

« Lemme », généralement renseignée, permet de repérer les éventuels homonymes (on n'affectera pas le nom « avions », de lemme « avion » à la même famille que le verbe « avions » de lemme « avoir »). Pour chaque mot affecté, le module indique, dans la fenêtre de gauche, les rangs d'apparition de ses occurrences dans les textes du corpus où il apparait. La fenêtre à droite permet de repérer, via leurs rangs d'apparition et le rang dans le texte des phrases auxquelles elles appartiennent, les occurrences des mots affectés au même synonyme.

Pour notre corpus, le mot « rêve » a été affecté au synonyme « sommeil ». Le module repère les trois occurrences de ce mot dans le corpus aux rangs 56, 64 et 180 du conte *Le rêve*. Il repère aussi, entre autres, l'unique occurrence du mot « endormir », lui aussi affecté à ce synonyme, comme 8^{e} occurrence du texte *Le communiqué*. En calculant le rapport entre le nombre d'occurrences des synonymes du mot et le nombre de mots de chaque texte, on remarque que le texte *Le rêve* est nettement plus « coloré » par ces synonymes que les autres (7/184=4% des mots contre respectivement 0.9% et 1.5% pour *Les chats* et *Le communiqué*).

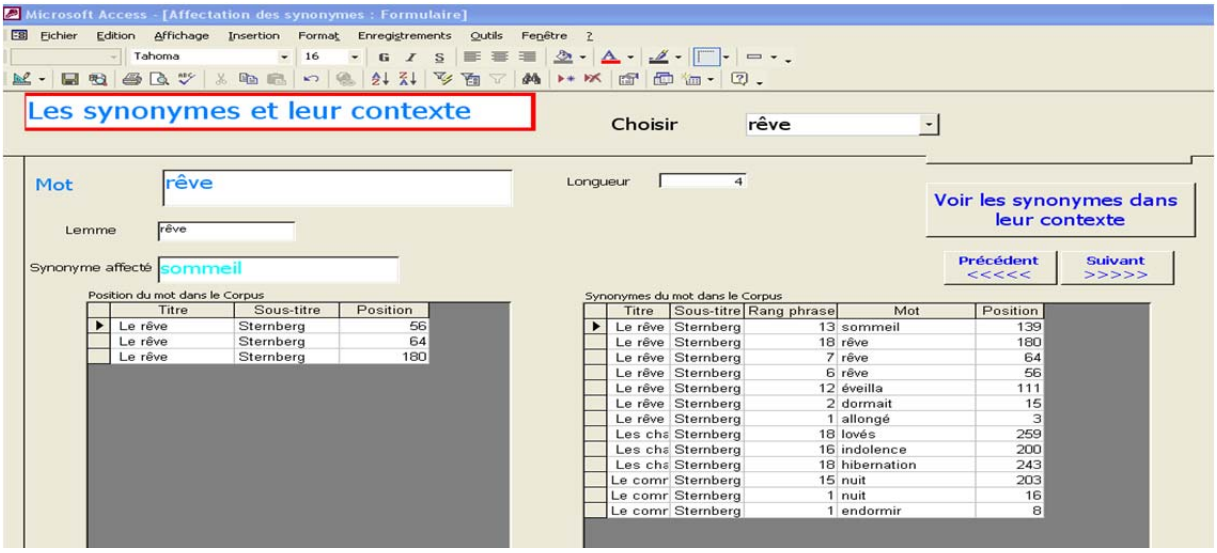


FIGURE 6 – *Les mots affectés au synonyme « sommeil » dans le corpus*

A travers le sous-menu « Voir les synonymes dans leur contexte », le module permet en outre d'étudier sommairement les contextes dans lesquels se trouvent les mots affectés à un même synonyme dans le corpus. Pour un mot donné, le module permet de visualiser la liste des contextes dans lesquels il est utilisé ainsi que les contextes « synonymisés » correspondants. Le contexte est ici constitué de la phrase dans laquelle se trouve le mot d'intérêt. Des informations comme le nombre d'occurrences du mot et de son synonyme dans la phrase, dans le texte et dans le corpus sont données. On peut faire défiler les phrases contenant le synonyme à l'aide des boutons « Suivant » et « Précédent ». Chaque phrase peut en outre être visualisée plus clairement à l'aide du bouton « Voir la phrase ».

Notons que ce module a une utilité multiple. D'une part, il permet de visualiser les contextes afin de vérifier la pertinence des synonymes (et, comme cas particulier, des lemmes ou des thèmes) affectés à un mot. Il permet par ailleurs de se faire une idée de l'importance et de la distribution d'un synonyme dans le corpus. Par ailleurs, il permet aussi l'étude du contexte d'un mot simple, rendue possible en affectant ce mot lui-même dans la fenêtre « Synonyme affecté ». Cette étape est alors appelée l'étude des concordances en analyse

textuelle et représente une phase importante de l'analyse, un mot tirant son sens de son « entourage », c'est-à-dire des mots auxquels il est associé.

Dans notre exemple, parmi les 3 occurrences du synonyme « sommeil » dans le texte *Les chats*, deux se trouvent dans la 18e phrase (une dans la 16e), c'est-à-dire dans la partie finale du texte qui en compte 23. Par ailleurs, ce qui n'apparaît pas ici dans la capture d'écran, on trouve aussi dans la 18e phrase deux mots affectés au synonyme « confort » (douillettement et bien-être), laissant apparaître un probable lien entre ces deux notions.

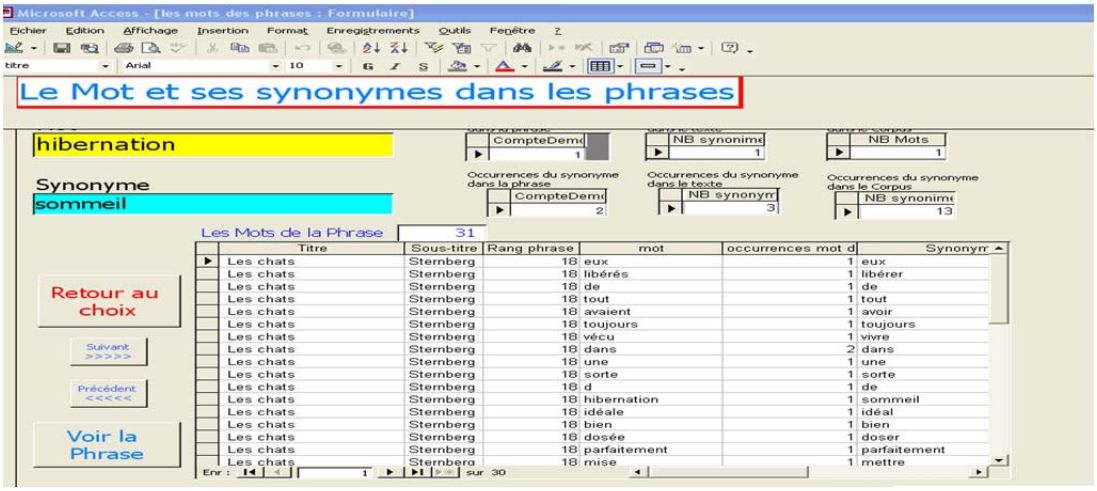


FIGURE 7 – *Contexte des mots de la famille « Sommeil » dans la 18e phrase du texte* Les chats

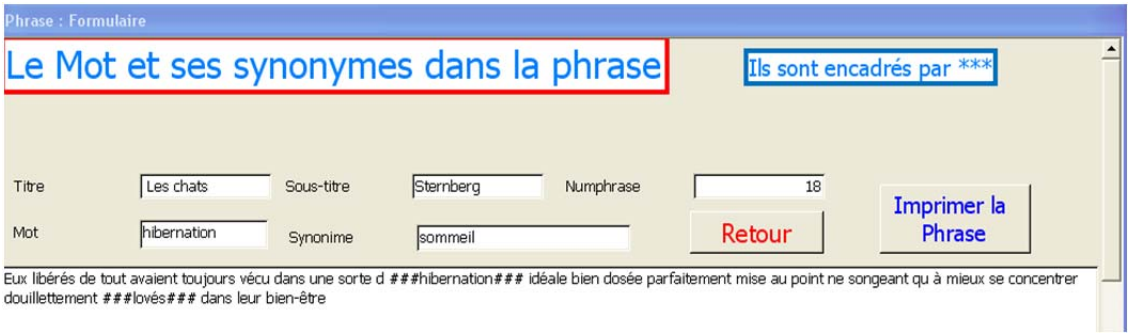


FIGURE 8 – *Visualisation de la 18e phrase du texte* Les chats

Mesure de l'adéquation d'un texte à une thématique

L'objectif du module, accessible par le menu « Thèmes du lexique », est de quantifier l'adéquation des idées des textes du corpus à certaines thématiques prédéfinies. Notons que thème revêt un sens plus large que synonyme, plusieurs synonymes pouvant être affectés à un même thème. La création d'une nouvelle thématique se fait en cliquant sur le bouton « Nouveau thème » et en entrant l'intitulé du thème dans la fenêtre « Thème ». Le thème ainsi créé est automatiquement numéroté et stocké dans une table où figurent l'ensemble des thèmes créés depuis la première utilisation du logiciel, que l'on peut faire défiler à l'aide des boutons « Précédent » et « Suivant ». A chaque thématique peut être associé un lexique de mots du dictionnaire que l'on définit dans la fenêtre « Mots du lexique » à l'aide du menu déroulant situé à gauche des cellules de la colonne « mot » (pour un thème déjà créé, les mots

affectés lors d'une utilisation antérieure du logiciel sont déjà présents dans la fenêtre). L'utilisateur, avec l'aide éventuelle d'un expert, doit définir deux poids pour chacun de ces mots. Ces poids, appelés force positive et force négative, mesurent respectivement l'adéquation et l'inadéquation du mot à la thématique étudiée. Le bouton « Voir la statistique » permet d'accéder à différents indicateurs aidant à appréhender l'intensité de la cohérence des textes avec les thématiques définies. Après avoir précisé le texte et la thématique d'intérêts, le module calcule les forces positives et négatives globales du texte pour cette thématique (somme des forces correspondantes des mots du texte associés à cette thématique), ainsi que la force globale (somme des valeurs absolues des forces globales négatives et positives). La puissance du lexique indique la somme des forces (positives, négatives, globales) des mots se trouvant dans le lexique de la thématique. Les ratios BRUT et CORRIGE donnent la force globale respectivement rapportée au nombre de mots du texte et à la puissance du lexique.

Dans notre corpus, nous avons étudié le thème de la mort. Nous avons affecté au lexique des mots du dictionnaire qui nous semblaient liés à ce thème et défini pour chacun d'eux des forces inférieures ou égales à 5 en valeur absolue. Ainsi, le mot « vivait » a une force positive de 0 et une force négative de 5, indiquant par là qu'il est complètement opposé à la notion de mort. N'ayant pas créé un lexique complet du thème (seulement 32 mots du dictionnaire y ont été affectés), les statistiques délivrées ici ne sont pas toutes pertinentes. On peut cependant constater que dans *Les chats*, bien que le thème de la mort soit présent, il existe une neutralité dans la coloration du texte par rapport à ce thème (la force positive est sensiblement égale à la force négative du texte sur ce thème).

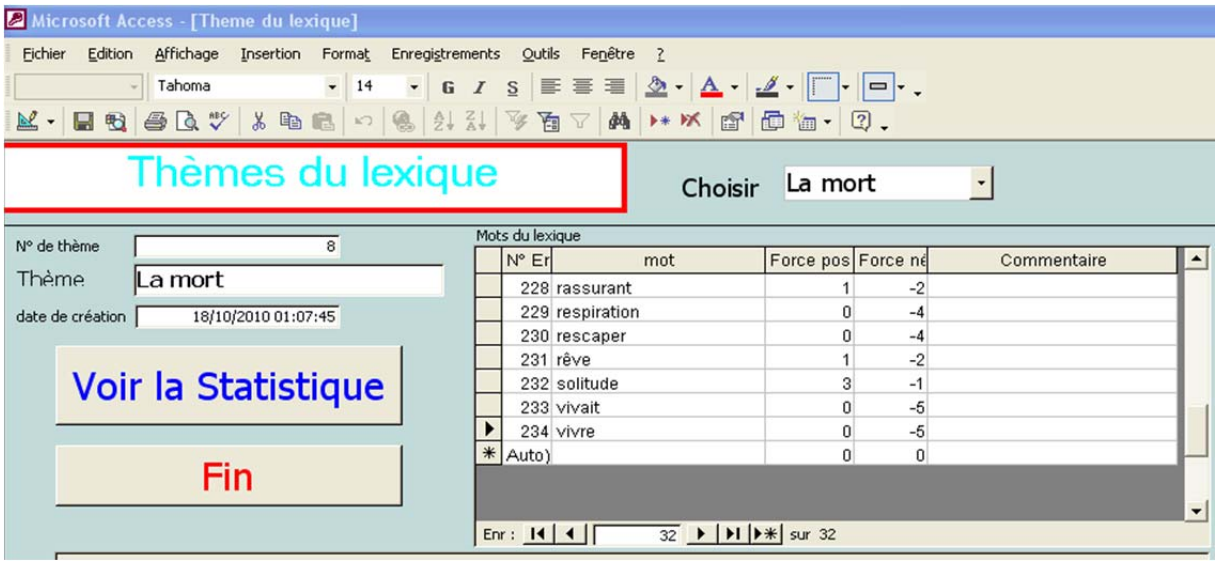


FIGURE 9 – *Création du lexique du thème « La mort »*

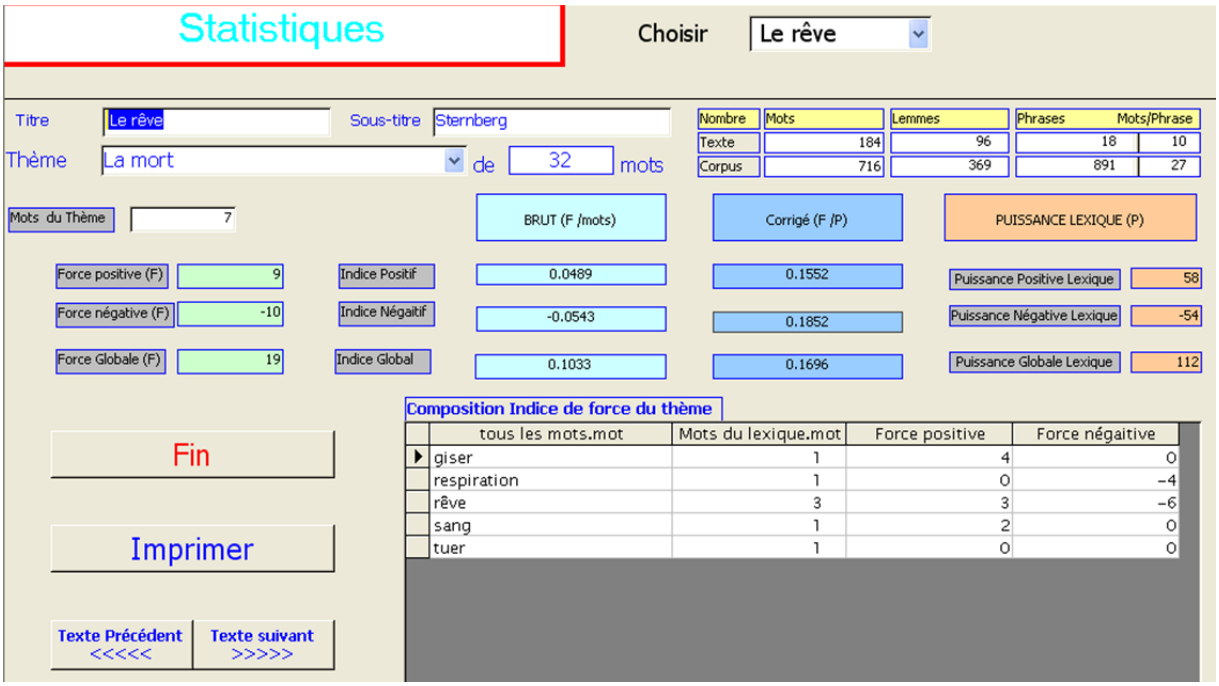


FIGURE 10 – *Pregnance du thème « La mort » dans le texte* Le rêve

3 Conclusion

La motivation préalable à la conception de cet outil était d'accompagner les étudiants dans la mise en application de leurs connaissances en analyse textuelle et en Gestion de Bases de Données. A posteriori, le bilan est positif : les étudiants réussissent à s'approprier le logiciel et à le faire évoluer au gré de leurs besoins. Ils sont en mesure de comprendre son architecture et de modifier ou créer des modules complémentaires ; dernièrement, un module a été élaboré sous ACCESS par un groupe d'étudiants. Il permet de tracer le diagramme de Pareto d'un texte, qui représente graphiquement la gamme des fréquences de ses formes.

Références

[1] Adam, J.-M. (2008), *La linguistique textuelle*, Armand Colin, Paris.

[2] De Saint Pol, T. (2003), *La statistique textuelle et ses logiciels*, http://www.melissa.ens-cachan.fr/IMG/ppt/statttext-2.ppt

[3] Gauzente, C. et D. Peyrat-Guillard (2007), *Analyse statistique de données textuelles en sciences de gestion*, ems, Paris.

[4] Jeandillou, J.-F. (2006), *L'analyse textuelle*, Armand Colin, Paris.

[5] Lebart, L. et A. Salem (1994), *L'analyse des données textuelles*, Dunod, Paris.

[6] Lundquist, L. (1983), *L'analyse textuelle. Méthode, Exercices*, Cedic, Paris